

Revisiting Differentially Private Regression: Lessons From Learning Theory and their Consequences

Xi Wu[†] Matthew Fredrikson[‡] Wentao Wu^{*} Somesh Jha[†] Jeffrey F. Naughton[†]

[†]University of Wisconsin-Madison, [‡]Carnegie Mellon University

^{*}Microsoft Research

{xiwu, wentaowu, jha, naughton}@cs.wisc.edu, mfredrik@cs.cmu.edu,

December 22, 2015

Abstract

Private regression has received attention from both database and security communities. Recent work by Fredrikson et al. (USENIX Security 2014) analyzed the functional mechanism (Zhang et al. VLDB 2012) for training linear regression models over medical data. Unfortunately, they found that model accuracy is already unacceptable with differential privacy when $\varepsilon = 5$. We address this issue, presenting an explicit connection between differential privacy and stable learning theory through which a substantially better privacy/utility tradeoff can be obtained. Perhaps more importantly, our theory reveals that the most basic mechanism in differential privacy, *output perturbation*, can be used to obtain a better tradeoff for *all* convex-Lipschitz-bounded learning tasks. Since output perturbation is simple to implement, it means that our approach is potentially widely applicable in practice. We go on to apply it on the same medical data as used by Fredrikson et al. Encouragingly, we achieve accurate models even for $\varepsilon = 0.1$. In the last part of this paper, we study the impact of our improved differentially private mechanisms on *model inversion attacks*, a privacy attack introduced by Fredrikson et al. We observe that the improved tradeoff makes the resulting differentially private model *more susceptible* to inversion attacks. We analyze this phenomenon formally.

1 Introduction

Differential-private data analytics has received considerable attention from the data management community [16, 19, 31, 32]. This paper focuses on regression models, which are widely used to extract valuable data patterns in “sensitive” domains. For example, in personalized medicine, linear regression models are commonly used to predict medication dosages [14] and other useful features [6, 21]. For such scenarios, individuals’ privacy has become a major concern and learning with differential privacy (DP) is particularly desirable.

Differentially private regression, and more generally private convex learning, has been intensively studied in recent years by researchers from the data management, security, and theory communities [3, 5, 7, 11, 15, 33]. One notable contribution is the *functional mechanism*, proposed by Zhang et al. [33], which is a practical mechanism for training differentially-private regression models. Several research groups [1, 29, 30] have adopted the functional mechanism as a basic building block in their study and initial empirical results are promising.

Our starting point is recent work by Fredrikson et al. [11], which used the functional mechanism to train differentially-private linear regression models to predict doses of a medicine called *warfarin* from patients’ genomic traits. Unfortunately, they found that the model accuracy was unacceptable with ε -DP — even for ε as high as 5. One of the main goals of this paper is to develop a suitable theory which allows mechanisms based on simple techniques such as *output perturbation* to obtain accurate, differentially-private models at reasonable private levels (e.g., $\varepsilon = 0.1$).

To this end, we start by presenting an explicit and precise connection between differential privacy and stability theory in computational learning. Specifically, in the setting of learning, we show that differential privacy is essentially another stability definition, yet it is so strong that it implies previous stability definitions in the learning literature. Combining this observation with some machinery from stable learning theory, we give an analysis showing that simple mechanisms such as output perturbation can learn *every* convex-Lipschitz-bounded problem with *strong* differential privacy guarantees.

Our analysis has several advantages over previous work. First, it *relaxes* the technical conditions required by Chaudhuri et al. [5]. In particular, we do not require the loss function to be smooth or differentiable. Second, our analysis reveals that output perturbation can be used to obtain a much better privacy/utility tradeoff than the functional mechanism. Under the same technical conditions, we achieve the *same tradeoff* between differential privacy and generalization error as that recently proved by Bassily et al. [3], while avoiding use of the exponential mechanism [22] and the sophisticated sampling sub-procedure used by Bassily et al. [3]. This makes our approach widely-applicable in practical settings.

We then apply our theory to *linear regression models*, for which we notice that the functional mechanism is the state-of-the-art approach being applied in the literature [1, 11, 29]. Our mechanisms are *regularized* versions of least squares optimization. In contrast to the functional mechanism, where regularization is a heuristic, regularization is crucial for our theoretical guarantee.

Regularization adds additional parameters that need to be picked carefully in order to maintain DP. A standard approach used in related contexts is to employ a private parameter tuning algorithm [5] to select a set of parameters that depends on the training data. We have implemented this approach to parameter tuning and refer to the mechanisms using it as the *privately-tuned mechanisms* in our empirical study. We also consider a different approach, based on our proof of generalization error, that picks the parameters in a *data independent* manner.

We evaluate both approaches using the same data and experimental methodology used by Fredrikson et al. [11]. The results are encouraging — both approaches produce substantially more accurate models than the functional mechanism of Zhang et al. [33]. Specifically, two tuned mechanisms obtain accurate models with ϵ -DP for $\epsilon = 0.3$. Interestingly, the data-independent mechanism significantly outperforms the tuned ones for small ϵ and obtains accurate models even for $\epsilon = 0.1$! Finally, using the same simulation approach described by Fredrikson et al. [11], we note that as compared to the functional mechanism, our methods induce significantly smaller risk of mortality, bleeding and stroke, especially when compared at small ϵ settings.

Our results described thus far are positive, and follow in the tradition of a large body of research on differential privacy, which seeks to improve utility/privacy tradeoffs for various problems of interest. However, Fredrikson et al. [11] did not only consider differential privacy, they also considered an attack they term *model inversion*. As a simple example, consider a machine learning model w that takes features x_1, \dots, x_d and produces a prediction y . A model-inversion (MI) attack takes input x_1, \dots, x_{d-1} and a value y' that is related to y , and tries to predict x_d (thus “inverting the model”). For example, in [11], the authors consider the case where x_d is a genetic marker, y is the warfarin dosage, and x_1, \dots, x_{d-1} are general background information such as height and weight. They used an MI attack to predict an individual’s *genetic markers* based on his or her *warfarin dosage*, thus violating that individual’s privacy.

In the final part of our paper, we consider the impact of our improved privacy-utility tradeoff on MI attacks. Here our results are less positive: in improving the privacy-utility tradeoff, we have increased the effectiveness of MI attacks. While unintended, upon deeper reflection this is not surprising: simply put, the improved privacy-utility tradeoff results in less noise added, and less noise added means the model is easier to invert. We formalize this discussion and prove that this phenomenon is quite general, and not an artifact of our approach or this specific learning task.

What this means for privacy is an open question. If one “only cares” about differential privacy, then the increased susceptibility to MI attacks is irrelevant. However, if one believes that MI attacks are significant (and anecdotal evidence suggests that some medical professionals are concerned about MI attacks), then the fact that improved differential privacy can mean worse MI exposure warrants further study. In this direction, our work indeed extends a long line of work that discusses the interaction of DP and *attribute*

privacy [18, 20, 23], and gives a realistic application where misconceptions about DP can lead to unwanted disclosure. It is our hope that our work might highlight this issue and stimulate more discussion.

Our technical contributions can be summarized as follows:

- We give an explicit connection between differential privacy and stability theory in machine learning. We show that differential privacy is essentially a strong stability notion that implies well-known previous notions.
- Moreover, combining this connection with some machinery from stability theory, we prove that the simple output perturbation mechanism can learn *every* convex-Lipschitz-bounded learning problems with strong differential privacy. Our analysis relaxes the technical conditions required by Chaudhuri et al. [5], achieves the same tradeoff between DP and generalization error as proved by Bassily et al. [3], and is much simpler than both analyses.
- Since output perturbation is one of the simplest mechanisms to implement, it means that our method is widely applicable in practice. We go on to apply the theory to linear regression. We present and analyze regularized variants of linear regression, and give a detailed description on how to privately select parameters for regularization.
- We perform a re-evaluation, using the same medical data set and experimental methodology as previous work [11], comparing the functional mechanism with our mechanisms to train linear regression models. Encouragingly, we observe a substantially better tradeoff between differential privacy and utility.
- We perform a re-evaluation of model inversion in the same medical data analysis setting, and demonstrate that as we improve differentially-private mechanisms, MI attacks become more problematic. We provide a theoretical explanation of this phenomenon.

The rest of the paper is organized as follows: In Section 2 we present background knowledge that will facilitate our discussion later. Next we present the connection between differential privacy and stability theory in Section 3, and discuss the applications in Section 4. In Section 5, we compare output perturbation and the functional mechanism with respect to the privacy-utility or model-inversion efficacy tradeoff. Finally, Section 6 discusses related work and in Section 7, we provide concluding remarks.

2 Preliminaries

In this section we present some background in three different areas: learning theory, convex optimization and differential privacy. We will only cover some basics of these areas to facilitate our discussion later. Readers are referred to [4], [10] and [24] for an in-depth introduction to these three topics, respectively.

2.1 Learning Theory

Let X be a feature space, Y be an output space, and $Z = X \times Y$ be a sample space. For example, Y is a set of labels for classification, or an interval in \mathbb{R} for regression. Let \mathcal{H} be a hypothesis space and ℓ be an *instance-wise* loss function, such that on an input hypothesis $w \in \mathcal{H}$ and a sample $z \in Z$, it gives a loss $\ell(w, z)$.

Let \mathcal{D} be a distribution over Z , the goal of learning is to find a hypothesis $w \in \mathcal{H}$ so as to minimize its *generalization error* (or *true loss*), which is defined to be $L_{\mathcal{D}}(w) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(w, z)]$. The minimum generalization error achievable over \mathcal{H} is denoted as $L_{\mathcal{D}}^* = \min_{w \in \mathcal{H}} L_{\mathcal{D}}(w)$. In learning, \mathcal{D} is *unknown* but we are given a training set $S = \{z_1, \dots, z_n\}$ drawn i.i.d. from \mathcal{D} . The *empirical loss function* is defined to be $L_S(w) = \frac{1}{n} \sum_{i=1}^n \ell(w, z_i)$.

We are ready to define learnability. Our definition follows Shalev-Shwartz et al. [26] which defines learnability in the Generalized Learning Setting considered by Haussler [13]. This definition also directly generalizes PAC learnability [27].

Definition 1 (Learnability). *A problem is called agnostically learnable with rate $\varepsilon(n, \delta) : \mathbb{N} \times (0, 1) \mapsto (0, 1)$ if there is a learning rule $A : Z^n \mapsto \mathcal{H}$ such that for any distribution \mathcal{D} over Z , given $n \in \mathbb{N}$ and $\delta \in (0, 1)$, with probability $1 - \delta$ over $S \sim \mathcal{D}^n$, $L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}^* + \varepsilon(n, \delta)$. Moreover, we say that the problem is agnostically learnable if for any $\delta \in (0, 1)$, $\varepsilon(n, \delta)$ vanishes to 0 as n tends to infinity.*

We stress that in this definition, the generalization error holds *universally* for *every* distribution \mathcal{D} on the data. Intuitively, this definition says that given a confidence parameter δ and a sample size n , the learned hypothesis $A(S)$ is $\varepsilon(n, \delta)$ close to the best achievable. Note that $\varepsilon(n, \delta)$ measures the rate we converge to the optimal. Throughout this paper, we will use the *generalization error* as the *utility measure* of a hypothesis w . Naturally, the utility of a hypothesis is high if its generalization error is small.

In the above, the learning rule A is *deterministic* in the sense that it maps a training set deterministically to a hypothesis in \mathcal{H} . We will also talk about *randomized* learning rules, which maps a training set to a distribution over \mathcal{H} . More formally, a randomized learning rule \tilde{A} takes the form $Z^n \mapsto \mathcal{D}(\mathcal{H})$, where $\mathcal{D}(\mathcal{H})$ is the set of probability distributions over \mathcal{H} . The empirical risk of \tilde{A} on a training item z is defined as $\ell(\tilde{A}(S), z) = \mathbb{E}_{w \sim \tilde{A}(S)}[\ell(w, z)]$. The empirical risk of \tilde{A} on a training set S is defined as $L_S(\tilde{A}(S)) = \mathbb{E}_{w \sim \tilde{A}(S)}[L_S(w)]$. Finally, we define the generalization error of \tilde{A} as $L_{\mathcal{D}}(\tilde{A}(S)) = \mathbb{E}_{w \sim \tilde{A}(S)}[L_{\mathcal{D}}(w)]$. In short, we take expectation over the randomness of \tilde{A} .

Stability Theory. Stability theory is a sub-theory in machine learning (see, for example, Chapter 13 of [24] for a gentle survey of this area). As we will see, more stable a learning rule is, better DP-utility tradeoff we can achieve in a private learning. To discuss stability, we need to define “change of input data set.” We will use the following definition.

Definition 2 (Replace-One Operation). *For a training set S , $i \in [n]$ and $z' \in Z$, we define $S^{(i, z')}$ to be the training set obtained by replacing z_i by z' . In other words,*

$$S = \{z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_n\},$$

$$S^{(i, z')} = \{z_1, \dots, z_{i-1}, z', z_{i+1}, \dots, z_n\}.$$

Moreover, we write $S^{(i)}$ instead of $S^{(i, z')}$ if z' is clear from the context.

Informally, a learning rule A is stable if $A(S)$ and $A(S^{(i)})$ are “close” to each other. There are many possible ways to formulate what does it mean by “close.” We will discuss the following definition, which is the *strongest* stability notion defined by Shalev-Shwartz et al. [26].

Definition 3 (Strongly-Uniform-RO Stability [26]). *A (possibly randomized) learning rule A is strongly-uniform-RO stable with rate $\varepsilon_{\text{stable}}(n)$, if for all training sets S of size n , for all $i \in [n]$, and all $z', \bar{z} \in Z$, it holds that $|\ell(A(S^{(i)}), \bar{z}) - \ell(A(S), \bar{z})| \leq \varepsilon_{\text{stable}}(n)$.*

Intuitively, this definition captures the following property of a good learning algorithm A : if one changes *any one training item* in the training set S to get S' , the two hypotheses computed by A from S and S' , namely $A(S)$ and $A(S')$, will be “close” to each other (here “close” means that for any instance \bar{z} sampled from \mathcal{D} , the loss of $A(S)$ and $A(S')$ on \bar{z} are close).

A fundamental result on learnability and stability, proved recently by Shalev-Shwartz et al. [26], states the following,

Theorem 4 ([26], informal). *Consider any learning problem in the generalized learning setting as proposed by Vapnik [28]. If the problem is learnable, then it can be learned by a (randomized) rule that is strongly-uniform-RO stable.*

Qualitatively, this theorem says that *learnability and stability are equivalent*. That is, every problem that is learnable can be learned *stably* (under strongly-uniform-RO stability). In the context of differential privacy, this indicates that for any learnable problem one might hope to achieve a good DP-utility tradeoff. However, *quantitatively* the situation is much more delicate: as we will see later, strongly-uniform-RO stability is somewhat too weak to lead to a differential privacy guarantee without additional assumptions.

2.2 Convex Optimization

We will need the following basic concepts from convex optimization. Let \mathcal{H} be a closed convex set equipped with a norm $\|\cdot\|$. A set \mathcal{H} is R -bounded if $\|x\| \leq R$ for any $x \in \mathcal{H}$. A function $f : \mathcal{H} \mapsto \mathbb{R}$ is *convex* if for every $u, v \in \mathcal{H}$, and $\alpha \in (0, 1)$ we have

$$f(\alpha u + (1 - \alpha)v) \leq \alpha f(u) + (1 - \alpha)f(v).$$

Moreover, f is λ -strongly convex if instead

$$f(\alpha u + (1 - \alpha)v) \leq \alpha f(u) + (1 - \alpha)f(v) - \frac{\lambda}{2}\alpha(1 - \alpha)\|u - v\|^2.$$

A function $f : \mathcal{H} \mapsto \mathbb{R}$ is ρ -Lipschitz if for any $u, v \in \mathcal{H}$, $|f(u) - f(v)| \leq \rho\|u - v\|$. Let \mathcal{H} be a closed convex set in \mathbb{R}^d . A *differentiable* function $f : \mathcal{H} \mapsto \mathbb{R}$ is β -smooth if its gradient ∇f is β -Lipschitz (note that ∇f is a d -dimensional vector valued function). That is, $\|\nabla f(u) - \nabla f(v)\| \leq \beta\|u - v\|$. We will use the following property of strongly convex functions. It says that any convex function can be turned into a strongly convex one by adding to it a strongly convex function.

Lemma 5 ([4]). *If f is λ -strongly convex and g is convex then $(f + g)$ is λ -strongly convex.*

2.3 Differential Privacy

Given two databases D, D' of size n , we say that they are *neighboring* if they differ in at most one tuple. We define bounded ε -differential privacy.

Definition 6 (Bounded Differential Privacy).

A mechanism \mathcal{M} is called bounded ε -differentially private, if for any neighboring D, D' , and any event $E \subseteq \text{Range}(\mathcal{M})$, $\Pr[\mathcal{M}(D) \in E] \leq e^\varepsilon \Pr[\mathcal{M}(D') \in E]$.

This is called “bounded” differential privacy because the guarantee is over databases of size n . Note that all previous work in differentially private learning is for bounded differential privacy [3, 5, 33]. Dwork, McSherry, Nissim and Smith [8] provide an output perturbation method for ensuring differential privacy. This method is based on estimating the “sensitivity” of a query, defined as follows.

Definition 7. *Let q be a query that maps a database to a vector in \mathbb{R}^d . The ℓ_2 -sensitivity of q is defined to be*

$$\Delta_2(q) = \max_{D \sim D'} \|q(D) - q(D')\|_2.$$

Note that we are using the 2-norm sensitivity instead of the usual 1-norm. We have the following theorem.

Theorem 8 ([8]). *Let q be a query that maps a database to a vector in \mathbb{R}^d . Then publishing $q(D) + \kappa$ where κ is sampled from the distribution with density*

$$p(\kappa) \propto \exp\left(-\frac{\varepsilon}{\Delta_2(q)}\|\kappa\|_2\right)$$

ensures ε -differential privacy.

Importantly, the ℓ_2 -norm of the noise vector, $\|\kappa\|_2$, is distributed according to a Gamma distribution $\Gamma\left(d, \frac{\Delta_2(q)}{\varepsilon}\right)$. We have the following fact about Gamma distributions:

Theorem 9 ([5]). *For the noise vector κ , we have that with probability at least $1 - \gamma$, $\|\kappa\|_2 \leq \frac{d \ln(d/\gamma) \Delta_2(q)}{\varepsilon}$.*

2.4 Our Setting

In this paper, we will work in the same setting as in [3], where we assume the instance-wise loss function ℓ to be convex and Lipschitz. As we will see later, both conditions are crucial for achieving a good tradeoff between privacy and utility. We stress that *both* the convexity and Lipschitzness conditions here are *only with respect to w* . In other words, we only require $\ell(w, z)$ to be convex and Lipschitz *in w* (for any fixed $z \sim \mathcal{D}$ at a time). Note also that, differing from Chaudhuri et al. [5], we do *not* need ℓ to be differentiable in w . Finally, we will focus on ε -differential privacy. Our results readily extend to (ε, δ) -differential privacy by using a different noise distribution (e.g. Gaussian distribution) in output perturbation.

3 Differential Privacy and Stability Theory

In this section we present our results on the connection between differential privacy and stability theory. Our technical results can be summarized as follows:

DP implies Strongly-Uniform-RO Stability. In Section 3.1, we show that, in the setting of learning, differential privacy is a strong stability notion that implies strongly-uniform-RO stability. Strongly-Uniform-RO stability is the strongest stability notion proposed by Shalev-Shwartz et al. [26] from learning theory.

Norm Stability implies DP. In Section 3.2, we give a stability notion, ℓ_2 -RO stability, that leads to differential privacy by injecting a small amount of noise. ℓ_2 -RO stability is used implicitly in [25] to prove learnability of convex-Lipschitz-bounded problems under the condition that the instance loss function is smooth. Our analysis removes this requirement.

Simpler Mechanism with the Same Generalization Error. In recent work, Bassily et al. [3] give tight bounds (for both training and generalization errors) for differentially privately learning convex-Lipschitz-bounded problems. Their mechanisms require the exponential mechanism and a sophisticated sampling subprocedure. We show in Section 3.3 that the elementary output perturbation mechanism presented in [5] can give the same tradeoff between differential privacy and generalization error (however with weaker training error) for *every* convex-Lipschitz-bounded learning problem. Our proof relaxes the technical requirements of [5] (smoothness), and is significantly simpler than both [3, 5].

3.1 Differential Privacy is a Stability Notion

If one writes out the definition of bounded differential privacy (Definition 6) in the language of learning, it becomes: a learning rule A is ε -differentially private if, for all training set S of size n , for all $i \in [n]$, and all $z' \in Z$, and *any event* E , it holds that $\Pr[A(S) \in E] \leq e^\varepsilon \Pr[A(S^{(i)}) \in E]$, where the probability is taken over the randomness of A . When we contrast this definition with strongly-uniform-RO stability (Definition 3), the only difference is that the latter considers a particular type of event, namely for $\bar{z} \in Z$, the magnitude of $\ell(\tilde{A}(S), \bar{z})$. At this point, it is somewhat clear that differential privacy is essentially (yet another) stability notion. Nevertheless, it is so strong that it implies strongly-uniform-RO stability, as shown in the following:

Proposition 10. *Suppose that $|\ell(\cdot, \cdot)| \leq B$. Let $\varepsilon > 0$ and A be a randomized learning rule. If A is ε -differentially private, then it is strongly-uniform-RO stable with rate $\varepsilon_{\text{stable}} \leq B(e^\varepsilon - 1)$. Specifically, for $\varepsilon \in (0, 1)$, this is approximately $B\varepsilon$.*

Two remarks are in order. First, this implication holds without assuming anything on the loss function ℓ except for boundedness. Second, one may note that the *converse* of this proposition, however, is not true in general. For example, consider the case where ℓ is a constant function and $A(S) = h_1 \neq h_2 = A(S^{(i)})$. Then A is strongly-uniform-RO stable (with rate 0!) yet it is clearly *not* differentially private. Moreover, this example indicates that, even if strongly-uniform-RO stability has been achieved, one cannot hope for differential privacy by adding a “small amount” of noise to the output of A . This is because h_1 and h_2 can be arbitrarily far away from each other so the sensitivity of A cannot be bounded. This motivates us to define another stability notion for the purpose of differential privacy.

3.2 Norm Stability and Noise for DP

In this section we present a different stability notion that does lead to differential privacy by injecting a small amount of noise. We then use some machinery from stability theory to quantify the amount of noise needed. Following our discussion above, a natural idea now is that $A(S)$ and $A(S^{(i)})$ shall be close *by themselves*, rather than being close *under the evaluation of some functions*. Because the output of A lies in \mathcal{H} , which is a normed space¹ as long as perturbation on \mathcal{D} , a “universal” notion for closeness is that $A(S)$ and $A(S^{(i)})$ are close in norm. This leads to the following definition.

Definition 11 (ℓ_2 -RO Stability). *A learning rule A is ℓ_2 -RO stable with rate $\varepsilon(n)$, if for any $S \sim \mathcal{D}^n$, $z' \sim \mathcal{D}$ and $i \in [n]$, $\|A(S^{(i)}) - A(S)\|_2 \leq \varepsilon(n)$.*

Astute readers may realize that this is nothing more than a *rephrasing of the ℓ_2 -sensitivity* of a query (Definition 7). Thus, in the spirit of the output perturbation method mentioned in Section 2.3, if one can bound ℓ_2 -RO stability, then we only need to inject a small amount of noise for differential privacy.

If A is ℓ_2 -RO stable, then adding a small amount of noise to its output ensures differential privacy, and thus strongly-uniform-RO stability. However, without the Lipschitz condition, the resulting hypothesis might be useless. This is because a small distance (in ℓ_2 -norm) to $A(S)$ could give significant change in loss. This presents a barrier for proving learnability for the private mechanism. Thus in the following, we will restrict ourselves back to the setting discussed in Section 2.4 where we assume that ℓ is convex and Lipschitz (in w).

We now move on to quantifying the amount of noise needed for differential privacy. Specifically, we will show that for strongly-convex learning tasks, the “scale of the noise” we need is roughly only $O_{d,\varepsilon}(1/n)$ where n is the training set size (the big- O notation hides a constant that depends on the number of features d , and the DP parameter ε). This means that as training set size increases, the noise we need vanishes to zero for a fixed model and ε -DP. By contrast, for the functional mechanism, the “scale of the noise” is $O_{d,\varepsilon}(1)$. The following two lemmas are due to Shalev-Shwartz et al. [25]. We include their proofs in the appendix for completeness.

Lemma 12 (Exchanging Lemma). *Let A be a learning rule such that $A(S) = \arg \min_w \vartheta_S(w)$, where $\vartheta_S(w) = L_S(w) + \varrho(w)$ and $\varrho(w)$ is a regularizer. For any $S \sim \mathcal{D}^n$, $i \in [n]$ and $z' \sim \mathcal{D}$,*

$$\vartheta_S(u) - \vartheta_S(v) \leq \frac{\ell(v, z') - \ell(u, z')}{n} + \frac{\ell(u, z_i) - \ell(v, z_i)}{n},$$

where $u = A(S^{(i)})$ and $v = A(S)$.

Intuitively, this lemma concerns about the behavior of a learning rule A on neighboring training sets. A is a regularized learning rule: Its objective function is in the form of empirical risk $L_S(w)$ plus regularization error $\varrho(w)$ ($\varrho(\cdot)$ is called a regularizer). More specifically, this lemma upper bounds *the difference between the objective values of u and v* , that is $\vartheta_S(u)$ and $\vartheta_S(v)$, in terms of instance losses on the specific two instances that get exchanged.

Recall that our goal is to upper bound $\|u - v\|$. The following lemma accomplishes this task by upper bounding the norm of the difference by the difference of the objective values of u and v .

Lemma 13. *Let A be a rule where $A(S) = \arg \min_w \vartheta_S(w)$ and $\vartheta_S(w)$ is λ -strongly convex in w . Then for any $S \sim \mathcal{D}^n$, $i \in [n]$ and $z' \sim \mathcal{D}$, $\frac{\lambda}{2}\|u - v\|^2 \leq \vartheta_S(u) - \vartheta_S(v)$, where $u = A(S^{(i)})$, $v = A(S)$.*

To see this Lemma, we note that v is a *minimizer* of ϑ_S by the definition of the learning rule. Therefore by the definition of strong convexity, we have that for any $\alpha > 0$,

$$\begin{aligned} \vartheta_S(v) &\leq \vartheta_S((1 - \alpha)v + \alpha u) \\ &\leq \alpha \vartheta(v) + (1 - \alpha) \vartheta(u) - \frac{\lambda}{2} \alpha (1 - \alpha) \|u - v\|^2 \end{aligned}$$

¹ For simplicity, $\|\cdot\|$ refers to ℓ_2 -norm in the rest of the paper. Our results are applicable to other settings as long as perturbation is properly defined on the normed space.

The lemma is then proved by rearranging and tending α to 1. Intuitively, this lemma says that as long as the objective function $\vartheta_S(w)$ of A is good (that is, strongly convex), then one can upper bound the difference between u and v in norm by the difference between the objective values of u and v . Combining these two lemmas, we can prove the following main theorem in this section.

Theorem 14. *Let A be a learning rule with a λ -strongly convex objective loss function $\vartheta_S(w) = L_S(w) + \varrho(w)$ where $\varrho(w)$ is a regularizer. Assume further that for any $z \in Z$, $\ell(\cdot, z)$ is ρ -Lipschitz. Then A is $\frac{4\rho}{\lambda n}$ ℓ_2 -RO stable.*

This theorem says that if both the instance loss function and the objective function are well behaved (ℓ is ρ -Lipschitz and ϑ is strongly convex), then the learning rule A is roughly $(1/n)$ -norm stable (in other words, the sensitivity is $1/n$, which vanishes to 0 as training set size n grows). In the case when ℓ is already λ -strongly convex, we do not need a regularizer so one can set $\varrho(w) = 0$, hence a natural algorithm to ensure differential privacy in this case is to directly perturb the empirical risk minimizer with noise calibrated to its norm stability. Our discussion so far thus leads to Algorithm 1

Input: Privacy budget: $\varepsilon_p > 0$. Training set $S = \{(x_i, y_i)\}_{i=1}^n$.

Output: A hypothesis w .

- 1 Solve the empirical risk minimization $\bar{w} = \arg \min_w L_S(w)$.
- 2 Draw a noise vector $\kappa \in \mathbb{R}^d$ according to a distribution with density function $p(\kappa) \propto \exp\left(-\frac{\lambda n \varepsilon_p \|\kappa\|_2}{4\rho}\right)$.
- 3 Output $\bar{w} + \kappa$.

Algorithm 1: Output Perturbation for strongly-convex loss function: $\ell(w, z)$ is λ -strongly convex and ρ -Lipschitz in w , for every $z \in Z$.

Theorem 15. *Let (Z, \mathcal{H}, ℓ) be a learning problem where ℓ is λ -strongly convex and ρ -Lipschitz. Then Algorithm 1 is ε -differentially private.*

To see this, we note that $L_S(w)$ is λ -strongly convex because ℓ is, so we can set the objective function $\vartheta(w) = L_S(w)$. $\vartheta(w)$ is λ -strongly convex and ρ -Lipschitz, so Theorem 14 bounds its ℓ_2 -norm stability. The proof is then completed by plugging the stability bound into Theorem 8.

If the loss function is only convex (instead of being strongly convex), then the idea is to use a strongly convex regularizer to make the objective function strongly convex. Specifically, we use the Tikhonov regularizer $\varrho(w) = \lambda\|w\|^2/2$ (indeed, any strongly convex regularizer applies). This gives Algorithm 2.

Input: Privacy budget: $\varepsilon_p > 0$. Regularization parameter: $\lambda > 0$. Boundedness parameter: $R > 0$.

Training data: $S = \{(x_i, y_i)\}_{i=1}^n$.

Output: A hypothesis w .

- 1 Solve the regularized empirical risk minimization problem $\bar{w} = \arg \min_{\|w\| \leq R} (L_S(w) + \frac{\lambda}{2}\|w\|^2)$.
- 2 Draw a noise vector $\kappa \in \mathbb{R}^d$ according to a distribution where $p(\kappa) \propto \exp\left(-\frac{\lambda n \varepsilon_p \|\kappa\|_2}{4(\rho + \lambda R)}\right)$.
- 3 Output $\bar{w} + \kappa$.

Algorithm 2: Output Perturbation for general-convex loss function: $\ell(w, z)$ is convex and ρ -Lipschitz in w , for every $z \in Z$.

Theorem 16. *Let (Z, \mathcal{H}, ℓ) be a learning problem where ℓ is convex and ρ -Lipschitz. Suppose further that the hypothesis space \mathcal{H} is R -bounded. Then Algorithm 2 is ε -differentially private.*

The easiest way to see this theorem is to define a new loss function $\bar{\ell}(w, z) = \ell(w, z) + (\lambda\|w\|^2)/2$. Then $\bar{\ell}$ is λ -strongly convex and $(\rho + \lambda R)$ -Lipschitz over \mathcal{H} , and the theorem directly follows from Theorem 15. We remark that Algorithm 2 and Theorem 16 can be strengthened based on Theorem 14 and specifically do not rely on the boundedness condition. However, since our analysis of generalization error for this case (in the next section) critically relies on the boundedness condition, the algorithm and its privacy guarantee are stated in the current form.

3.3 Generalization Error

Until now, we have only talked about ensuring differential privacy with a small amount of noise, and have not said anything about whether this small amount of noise will lead to a model with “good utility”, which is usually measured by *generalization error* in learning theory. We accomplish this in this section. At a high level, we will show that for strongly convex learning tasks, output perturbation produces hypotheses that are roughly $(1/n)$ -away from the optimal. For general convex learning tasks, this degrades to roughly $1/\sqrt{n}$.

For notational convenience, throughout this section we let A denote a deterministic learning mechanism, and \tilde{A} be its output perturbation counterpart. We begin with a general lemma.

Lemma 17. *Suppose that for any $z \sim \mathcal{D}$, $\ell(w, z)$ is ρ -Lipschitz in w . If with probability at least $1 - \gamma$ over $w \sim \tilde{A}(S)$,*

$$\|w - A(S)\|_2 \leq \kappa(n, \gamma),^2$$

then with probability at least $1 - \gamma$ over $w \sim \tilde{A}(S)$,

$$|L_{\mathcal{D}}(w) - L_{\mathcal{D}}(A(S))| \leq \rho\kappa(n, \gamma).$$

This lemma translates the closeness between two hypotheses *in norm* to the closeness in *generalization error*. More specifically, note that the randomized learning rule \tilde{A} induces a *distribution* $\tilde{A}(S)$ over the hypothesis space. This lemma says as long as a hypothesis sampled from $\tilde{A}(S)$ is close to $A(S)$ (a single hypothesis) *in norm*, then these two hypotheses are close in their generalization error. Note that this closeness is controlled by the Lipschitz constant of the instance loss function ℓ .

Combing this lemma with Theorem 14 from the last section, which bounds the norm stability, we have the following general theorem upper bounding the generalization error of our method.

Theorem 18. *Let (\mathcal{H}, Z, ℓ) be a learning problem that is agnostically learnable by a deterministic learning algorithm A with rate $\varepsilon(n, \delta)$. Suppose that $\ell(w, z)$ is ρ -Lipschitz in w for any $z \in Z$. Let \mathcal{D} be a distribution over Z . Finally, suppose that for any $S \sim \mathcal{D}^n$, with probability at least $1 - \gamma$ over $w \sim \tilde{A}(S)$,*

$$\|w - A(S)\|_2 \leq \kappa(n, \gamma).$$

Then with probability at least $1 - \delta - \gamma$ over $S \sim \mathcal{D}^n$ and $w \sim \tilde{A}(S)$, we have $L_{\mathcal{D}}(w) - L_{\mathcal{D}}^ \leq \varepsilon(n, \delta) + \rho\kappa(n, \gamma)$.*

In the rest of this section we give concrete bounds for different types of instance loss function.

Strongly Convex Loss. We now bound generalization error of our method for the case where the instance loss function ℓ is strongly convex. We will use the following theorem from stability theory:

Theorem 19 (Theorem 6, [25]). *Consider a learning problem such that $\ell(w, z)$ is λ -strongly convex and ρ -Lipschitz in w . Then for any distribution \mathcal{D} over Z and any $\delta > 0$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,*

$$L_{\mathcal{D}}(\text{ERM}(S)) - L_{\mathcal{D}}^* \leq \frac{4\rho^2}{\delta\lambda n}.$$

where $\text{ERM}(S)$ is the empirical risk minimizer, i.e. $\text{ERM}(S) = \arg \min_{w \in \mathcal{H}} L_S(w)$.

We are ready to prove the following theorem on generalization error. Our bound matches the bound obtained by Bassily et al. for the same setting (Theorem F.2, [3]).

Theorem 20. *Consider a learning problem such that $\ell(w, z)$ is λ -strongly convex and ρ -Lipschitz in w . Let $\varepsilon_p > 0$ be a privacy parameter for differential privacy. Let \tilde{A} denote Algorithm 1. Then for any $\delta \in (0, 1)$, with probability $1 - \delta$ over $S \sim \mathcal{D}^n$ and $w \sim \tilde{A}(S)$,*

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}^* \leq O\left(\frac{\rho^2 d \ln(d/\delta)}{\lambda n \delta \varepsilon_p}\right).$$

²We abuse the notation κ to remind readers that this quantity is related to the noise vector.

Basically, this theorem says that if the instance loss function is strongly convex, then with high probability, the generalization error of a hypothesis sampled using our method is only roughly $1/n$ -away from the “optimal” ($L_{\mathcal{D}}^*$).

General Convex Loss. We now consider the general case where ℓ is only convex. We have the following theorem on the generalization error, which matches the bound obtained in [3] (Theorem F.3),

Theorem 21. *Consider a convex, ρ -Lipschitz learning problem that is also R -bounded. Let $\varepsilon_p > 0$ be a privacy parameter for differential privacy. Let \tilde{A} denote Algorithm 2. Then for any $\delta \in (0, 1)$, with probability $1 - \delta$ over $S \sim \mathcal{D}^n$ and $w \sim \tilde{A}(S)$,*

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}^* \leq O\left(\frac{\rho R \sqrt{d \ln(d/\delta)}}{\sqrt{n\delta\varepsilon_p}}\right).$$

Note that for convex loss functions (instead of strongly convex ones), we are only roughly $1/\sqrt{n}$ -away from the optimal.

We notice that regularization plays a vital role in this theoretical guarantee. Indeed, the key to prove Theorem 21 is the use of the Tikhonov regularizer ($\lambda\|w\|^2/2$) to gain *stability*. By contrast, regularization is only used as a heuristic in the analysis of the functional mechanism [33].

4 Applications

In this section we give concrete applications of the theory developed so far. Along the way, we compare our method with previous ones. To begin with, Chaudhuri et al. [5] have given an output perturbation mechanism that is essentially the same as our Algorithm 2. Moreover, they give an *objective perturbation* algorithm. The main difference is that the noise is added to the objective function, instead of the output. Algorithm 3 describes their objective perturbation algorithm. We refer interested readers to their work for more details.

<p>Input: Privacy budget: $\varepsilon_p > 0$. Parameters $\lambda, c > 0$. Boundedness parameter: $R > 0$. Training data: $S = \{(x_i, y_i)\}_{i=1}^n$.</p> <p>Output: A hypothesis w.</p> <ol style="list-style-type: none"> 1 Let $\varepsilon'_p = \varepsilon_p - \log(1 + \frac{2c}{n\lambda} + \frac{c^2}{n^2\lambda^2})$. 2 If $\varepsilon'_p > 0$, then $\Delta = 0$, else $\Delta = \frac{c}{n(e^{\varepsilon'_p/4}) - 1} - \lambda$, and $\varepsilon'_p = \varepsilon_p/2$. 3 Put $\beta = \varepsilon_p/2$ and draw noise vector ν such that $p(\nu) \propto e^{-\beta\ \nu\ }$. 4 Output $\arg \min_{\ w\ \leq R} \left\{ L_S(w) + \frac{\nu^T w}{n} + \frac{\lambda\ w\ ^2}{2} \right\}$.
--

Algorithm 3: Objective Perturbation of Chaudhuri et al.

Our work differs in analysis and applicability. Specifically, to get their claimed generalization bounds Chaudhuri et al. requires differentiability and the Lipschitz *derivative* for their output perturbation, and requires twice differentiability with bounded derivatives for their objective perturbation (see Theorem 6 and Theorem 9 in [5]). Our analysis removes all these differentiability conditions, and demonstrates that *output perturbation* works well for *all* convex-Lipschitz-bounded learning problems. This is by far the largest class of convex learning problems that are known to be learnable (see [24]). Due to lack of space, in the following we only give two important examples, Support Vector Machine and Logistic Regression. As we will see, our analysis enables us to train an SVM *without* approximating the loss function using a smooth loss function that gives higher or equal loss point-wisely.

Support Vector Machine (SVM). In SVM the instance-loss function is the so called *hinge loss function*. Specifically, given hypothesis space $\mathcal{H} \subseteq \mathbb{R}^d$, feature space $X \subseteq \mathbb{R}^d$, and output space $Y = \{0, 1\}$, then for $(x, y) \in X \times Y$, the hinge loss is defined as

$$\ell^{\text{hinge}}(w, x, y) = \max\{0, y(1 - \langle w, x \rangle)\}.$$

In the common setting where X is scaled to lie within the unit ball, it is straightforward to verify that ℓ^{hinge} is convex 1-Lipschitz. Therefore our Algorithm 2 and Theorem 21 directly apply to give differential privacy with a small generalization error. We note that, however, hinge loss is *not* differentiable. Therefore Chaudhuri et al.’s analysis for output perturbation does not directly apply. Indeed, in order to use their method, they need to use a smooth loss function to approximate the hinge loss by giving higher or equal loss point-wisely.

Logistic Regression. In the Logistic Regression we have $\mathcal{H} \subseteq \mathbb{R}^d$, $X \subseteq \mathbb{R}^d$ and $Y = \{\pm 1\}$. Then for $y \in Y$ and $x \in X$, the loss of w on (x, y) is defined to be

$$\ell^{\log}(w, x, y) = \log(1 + e^{-y\langle w, x \rangle}).$$

It is not hard to verify that ℓ^{\log} is convex and differentiable with bounded derivatives. Therefore both our analysis and Chaudhuri et al.’s analysis directly apply. In particular, Chaudhuri et al. [5] has done extensive experiments evaluating output perturbation for Logistic Regression on standard datasets.

4.1 On Setting Parameters, Generalization Error and Implementation

We note that, very recently, Bassily et al. [3] have given better *training error* under the *same* technical conditions as ours (the generalization error is the same). Next we compare with theirs.

On one hand, we observe that [3, 5], as well as our work, are all based on regularized learning, and the regularization parameter λ is left unspecified. Because the requirement of differential privacy, one cannot naively run the regularized learning on multiple training sets and then pick the best parameters: Our analysis of DP is with respect to the training data only; however, if the parameters we pick depend on the data, then they may carry sensitive information of individuals in the data, which renders the subsequent DP analysis invalid. A standard way to tackle this issue is to employ private parameter tuning, which acts as a *wrapper procedure*, that invokes the regularized learning procedure as a black box, and pick parameters differentially privately for the regularized learning. In this paper we also consider a *data independent* approach based on our analysis.

On the other hand, for fixed λ the regularized learning procedure of Bassily et al. [3] achieves better training error (the generalization error remains the same). However, their mechanisms are substantially more complicated. More specifically, their mechanism requires a sophisticated sampling procedure that runs in time $O(n^3)$, where n is the training set size. For reasonably large data sets this can be prohibitive. An even more serious concern is some practical challenges in implementing this sampling procedure, which dates back to work of Applegate and Kannan [2]. Indeed, we noticed that similar concern in implementing related sampling procedures have also been raised by Hardt [12]. Given these concerns, we will not empirically compare with Bassily et al.’s mechanisms.

4.2 Linear Regression

In the rest of this paper, we specialize the general algorithms to linear regression, which is a common regression task and is the model which Fredrikson et al. [11] constructed from a medical data set. We will evaluate the effectiveness of our private linear regression mechanisms presented here in an empirical study later.

Technically, let d be the number of features, $\mathcal{H} \subseteq \mathbb{R}^d$, $X \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}$. Linear regression uses the so called “*squared loss*” instance-loss function, $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$. Note that ℓ is convex in w for any fixed (x, y) . Unfortunately, it is known in the learning theory that the learning task with respect to ℓ (that is, to minimize $\mathbb{E}_{\mathcal{D}}[\ell(w, \mathcal{D})]$ over $w \in \mathcal{H}$) is not learnable over \mathbb{R}^d (see Example 12.8, Shalev-Shwartz and Ben-David [24]). To gain theoretical tractability, we thus restrict the hypothesis space \mathcal{H} to be R -bounded. With this restriction, ℓ is then $(2R + 2)$ -Lipschitz over \mathcal{H} . Thus we can specialize Algorithm 2 to get `rls-out` described in Algorithm 4.

Note that `rls-out` has two unspecified parameters λ and R . We use two approaches to tackle the issue of picking these parameter privately: The first is to use a computation based on our analysis to pick parameters

Input: Privacy budget: $\varepsilon_p > 0$. Regularization parameter: $\lambda > 0$. Boundedness parameter: $R > 0$.
Training data: $S = \{(x_i, y_i)\}_{i=1}^n$, where $\|x_i\| \leq 1$ and $|y_i| \leq 1$.
Output: A hypothesis w .

- 1 Solve the regularized least mean square problem

$$\bar{w} = \arg \min_{\|w\| \leq R} \left(L_S(w) + \frac{\lambda}{2} \|w\|^2 \right).$$

- 2 Draw a noise vector $\kappa \in \mathbb{R}^d$ according to a distribution with the following density function $p(\kappa)$,

$$p(\kappa) \propto \exp \left(-\frac{\lambda n \varepsilon \|\kappa\|_2}{12R + 8} \right).$$

- 3 Output $\bar{w} + \kappa$.

Algorithm 4: rls-out

in a data *independent* manner. The second approach uses a private parameter tuning algorithm (Section 6, [5]). We are ready to describe all the private linear regression mechanisms considered in this paper and explain on how to choose parameters. In the following, similar to previous work [5, 33], we assume that the training data is scaled so that they lie in the unit ball.

Oracle Output Perturbation. This is the *idealized* version of the output perturbation mechanism, where we exhaustively search for the best parameters using the training data and then use them in as if they were “constants”. We consider this variant because it shows the best possible result one can obtain using output perturbation.

We search R exhaustively within the space $\{0.25, 0.5, 1, 2\}$. To search for a good λ , we consider an arithmetic progression starting at initial value $(d/n\varepsilon'_p)^{1/2}$ and ending at $(d/n\varepsilon_p^*)^{1/2}$, where $\varepsilon'_p = 100$ and $\varepsilon_p^* = 0.1$. The start and end values are determined by a computation based on our proof of Theorem 21. For our dataset on warfarin dosage, $n \approx 3000$ and $d = 14$. This gives the approximate range $[0.007, 0.2]$. We use a slightly larger range $[0.001, 0.5]$.

Data-Independent Output Perturbation. Based on the proof of Theorem 21, We determine λ, R through a computation so that they are independent of the training data.

We put $R = 1$. This is because we scale the data to be within the unit ball, and the coefficients of the linear regression model are all small when trained over the scaled data. For λ , we pick it using our theoretical analysis for output perturbation. Specifically, Theorem 21 indicates that λ is approximately $\sqrt{d/n\varepsilon_p}$.

Privately-Tuned Output/Objective Perturbation. Chaudhuri et al. [5] use private parameter tuning to pick data-dependent parameters for both of their output/objective perturbation. Fortunately, the private tuning algorithm only makes black-box use of the output/objective perturbation. Algorithm 5 describes the private-tuning framework. Given ε_p , the tuning algorithm ensures ε_p -DP.

Let \mathcal{R} be a set of l choices of R and Λ be a set of k choices of λ . Let $m = l \cdot k$, and suppose the pairs of parameters can be listed as $\{(R_1, \lambda_1), \dots, (R_m, \lambda_m)\}$. The more settings of parameters to try, larger the m is, and so we have smaller chunks for training, thus more entropy in the probability in picking the final hypothesis. Therefore, we want to have a small set of good parameters. We put Λ as a geometric progression of ratio 2 that starts with 0.002 and ends with $0.256 = 0.002 \cdot 2^7$, so $k = 8$. For R , note that while we may want larger radius for the purpose of learning, the probability we assign to each w_i decays quadratically in R . We thus try two sets $\mathcal{R}_1 = \{.25, .5, 1\}$, and $\mathcal{R}_2 = \{.5, 1, 2\}$, so $l = 3$. For \mathcal{R}_1 , the probability sampling w_i becomes $\exp(-L_{S'}(w_i)\varepsilon_p/2)$. For \mathcal{R}_2 , this is $\exp(-L_{S'}(w_i)\varepsilon_p/8)$.

Input: Privacy budget: $\varepsilon_p > 0$.

Training data: $S = \{(x_i, y_i)\}_{i=1}^n$, where $\|x_i\| \leq 1$, $|y_i| \leq 1$.

Parameter space: $\mathcal{R} = \{R_1, \dots, R_l\}$, where $\max_{1 \leq i \leq l} R_i \leq R$, $\Lambda = \{\lambda_1, \dots, \lambda_k\}$. Let $\mathcal{R} \times \Lambda = \{(R_i, \lambda_i)\}_{i=1}^m$

Output: A hypothesis w .

- 1 Divide the training data into $m + 1$ chunks, S_1, \dots, S_m, S' . S_1, \dots, S_m are used for training, and S' is used for validation.
- 2 For each $i = 1, 2, \dots, m$, apply a privacy-preserving algorithm to train w_i (for example output or objective perturbation) with parameters $\varepsilon_p, \lambda_i, R_i, S_i$. Evaluate w_i on S' to get utility $u_{S'}(w_i) = -L_{S'}(w_i)$.
- 3 Pick a w_i in $\{w_1, \dots, w_m\}$ using the exponential mechanism with privacy budget ε_p and utility function $u_{S'}(w_i)$. Note that the sensitivity of u is

$$\Delta(u) = \max_{w \in \mathcal{H}} \max_{S, i, z'} |u_S(w) - u_{S^{(i)}}(w)| \leq R^2.$$

Thus this amounts to sampling w_i with probability

$$p(w_i) \propto \exp\left(-\frac{L_{S'}(w_i)\varepsilon_p}{2R^2}\right).$$

Algorithm 5: Parameter Tuning Algorithm: it accesses a privacy-preserving training algorithm as a black box.

5 Empirical Study

In this section we evaluate our mechanisms on the warfarin-dosing dataset used by Fredrikson et al. [11]. We are interested in the following questions:

1. How does the accuracy of models produced using our mechanisms compare to the functional mechanism?
2. What is the impact of improved DP-utility tradeoff on MI attacks?

In summary, we found that: 1) our mechanisms provide better accuracy than the functional mechanism, for a given ϵ setting, and 2) this improvement in model accuracy makes MI attacks more effective. Specifically, we obtain accurate models with ϵ -DP even for $\epsilon = 0.1$, whereas the functional mechanism does not provide comparable models until $\epsilon \geq 5$.

Section 5.1 describes our experimental methodology, and Section 5.2 presents our results on the relationship between ϵ and model accuracy. In Section 5.3, we present empirical evidence and a formal analysis of our observation that better differentially-private mechanisms lead to more effective MI attacks.

5.1 Methodology

We used the same methodology, data, and training/validation split as the experiments discussed in Fredrikson et al. [11]. The data was collected by the International Warfarin Pharmacogenetics Consortium (IWPC), and contains information pertaining to the age, height, weight, race, partial medical history, and two genomic SNPs: VKORC1 and CYP2C9. The outcome variable corresponds to the stable therapeutic dose of warfarin, defined as the steady-state dose that led to stable anticoagulation levels. At the time of collection, this was the most expansive database of information relevant to pharmacogenomic warfarin dosing. For more information about the data and how it was pre-processed, we refer the reader to the original IWPC paper [14] and the paper by Fredrikson et al. [11].

Our experiments examine linear warfarin dosing models trained on this data, either using differentially-private regression mechanisms or standard linear regression, and seek to characterize the relationship between the following quantities:

Privacy budget. This is given by the parameter ϵ , and controls the “strength” of the guarantee conferred by differential privacy. Smaller privacy budgets correspond to stronger guarantees.

Model accuracy/utility. We measure this as the average mean-squared error between the model’s predicted warfarin dose, and the ground truth given in the validation set. To obtain a more realistic measure of utility, we use the approach described by Fredrikson et al. [11] to simulate the physiological response of patients who are given warfarin dosages according to the model, and estimate the likelihood of adverse health events (stroke, hemorrhage, and mortality).

Model invertability. This is measured by the success rate of the model inversion algorithm described by Fredrikson et al. [11], when predicting VKORC1 for each patient in the validation set. We focused on VKORC1 rather than CYP2C9, as Fredrikson et al. showed that the results for this SNP more clearly illustrate the trade-off between privacy and utility.

The rest of this section describes our results for each quantity.

5.2 Model Accuracy

Figure 1 compares Functional Mechanism with all the private output perturbation mechanisms, which includes Tuned Output Perturbation (Out), Data-Independent Output Perturbation (Data Independent), and Oracle Output Perturbation (Oracle). We also include the model accuracy of the non-private algorithm (Non-Private). We observe that all the output perturbation give much better model accuracy compared to the functional mechanism, especially for small ϵ . Specifically, Tuned Output Perturbation obtains accurate models at $\epsilon = 0.3$. Data-Independent and the Oracle Mechanisms give much the same model accuracy, and provide accurate models even for $\epsilon = 0.1$. In particular, the accuracy is very close to the models produced by the non-private algorithm.

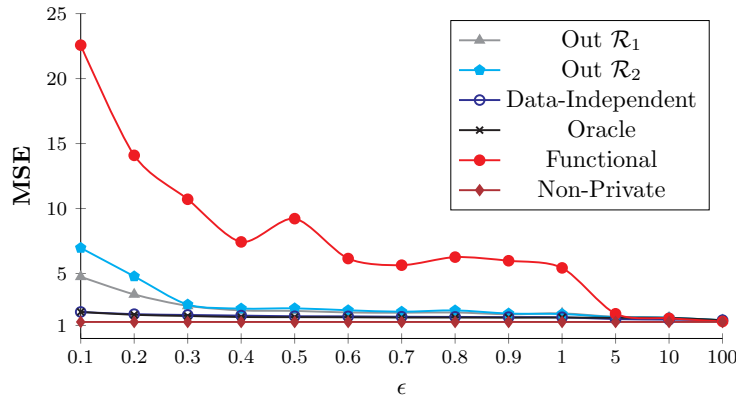


Figure 1: Model Accuracy: compare the functional mechanism with our output perturbation method, as well as the non-private mechanism. \mathcal{R}_1 stands for the Private-Tuned mechanism where we try the bounded hypothesis space with radius R in $\{.25, .5, 1\}$. \mathcal{R}_2 stands for the same mechanism with radius in $\{.5, 1, 2\}$. Out indicates Privately Tuned Output Perturbation methods. Oracle (Data-Independent) stands for the Oracle (Data-Independent) Output Perturbation.

Figure 2 further compares the performance of Data-Independent and Oracle mechanisms and demonstrate their closeness. For the entire parameter range considered, the maximum MSE gap we observed is only 0.1. Figure 3 further compares the risk of mortality, hemorrhage, and stroke using Functional Mechanism, Tuned Output Perturbation and Data-Independent Output Perturbation. unsurprisingly, Data-Independent Output Perturbation gives the best result, and in particular much smaller risk than Functional Mechanism.

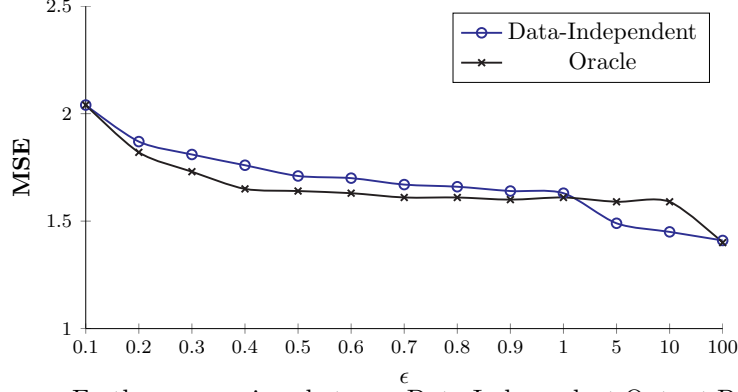


Figure 2: Model Accuracy: Further comparison between Data-Independent Output Perturbation and Oracle Output Perturbation. The maximum MSE gap we observed is only 0.1.

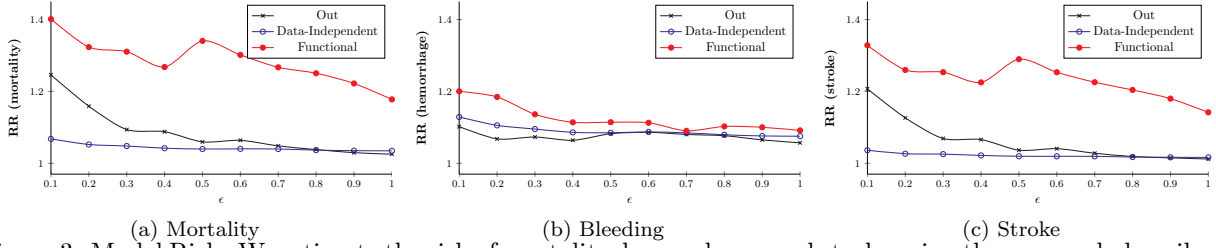


Figure 3: Model Risk: We estimate the risk of mortality, hemorrhage, and stroke using the approach described by Fredrikson et al.

Comparison with Other Private Mechanisms. We also compare model accuracy of our method with two other private algorithms.

Projected Histogram. We notice that in the previous work of Fredrikson et al. [11], they have implemented private projected histogram mechanism and compared it with the functional mechanism on linear regression. Specifically, as Figure 6 (Section 5.2) of their paper shows, the projected-histogram algorithm indeed has similar model accuracy compared to the functional mechanism, which is much worse than our Data-Independent algorithm (of which the accuracy is close to that of the non-private algorithm).

Objective Perturbation. We have so far mainly compared with the function mechanism, which we believe is the most important task, because the functional mechanism has become the recognized state of the art for training regression models and has been adopted by many research teams, including [1, 29, 30]. On the other hand, we notice that under very restricted conditions, it is known that Chaudhuri et al.’s objective perturbation method [5] can provide very good model accuracy. Interestingly, because we impose boundedness condition for linear regression, the technical conditions of objective perturbation are satisfied. Therefore we also compare it with our method. Encouragingly and perhaps somewhat surprisingly, while our method is much more widely applicable (see discussion in Section 4), our experiments show that our general algorithm performs as well as objective perturbation. Specifically, Figure 4 compares the model accuracy of these three methods, and the maximal gap we observed is only 0.1!

5.3 Model Inversion

In this section, we examine the impact of the increased model utility on *model inversion* (MI), a privacy attack first raised by Fredrikson et al. [11]. Improving model utility for a given ϵ is a theme shared by nearly all previous work on differential privacy. This is a sensible goal, because utility has no direct bearing on

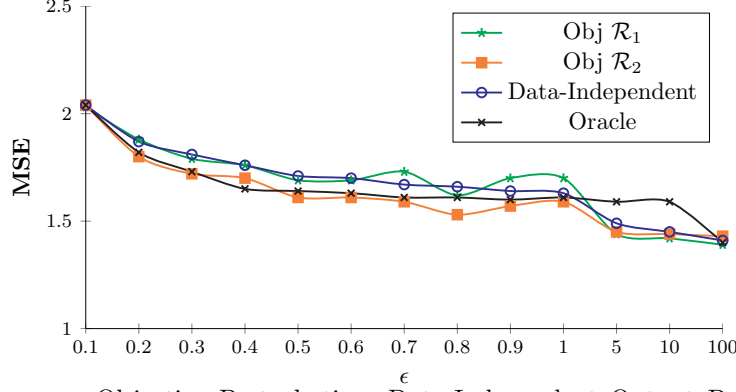


Figure 4: Model Accuracy: Objective Perturbation, Data-Independent Output Perturbation and Oracle Output Perturbation. \mathcal{R}_1 stands for the Private-Tuned mechanism where we try the bounded hypothesis space with radius R in $\{.25, .5, 1\}$. \mathcal{R}_2 stands for the same mechanism with radius in $\{.5, 1, 2\}$.

the privacy guarantee provided by *differential-privacy*—two models can differ significantly on the level of utility they provide, while still conferring the same level of differential privacy. However, we show that MI is orthogonal to differential privacy in this sense, because the improved utility offered by our mechanisms leads to more successful MI attacks.

Better DP mechanisms, more effective MI attacks. Figure 5 compares MI accuracy of all the private mechanisms. For all these mechanisms, we see that mechanisms with better DP-utility tradeoff also has higher MI accuracy. Specifically, For Oracle Output Perturbation at $\varepsilon = 0.2$, we see a significant increase in MI accuracy: 45% for Oracle compared to 35% for the functional mechanism. Meanwhile, the utility of our mechanism is much better, with mean squared error 1.82 compared to the functional mechanism’s 22.57. This phenomenon holds for larger ε , although the magnitude of the differences gradually shrink.

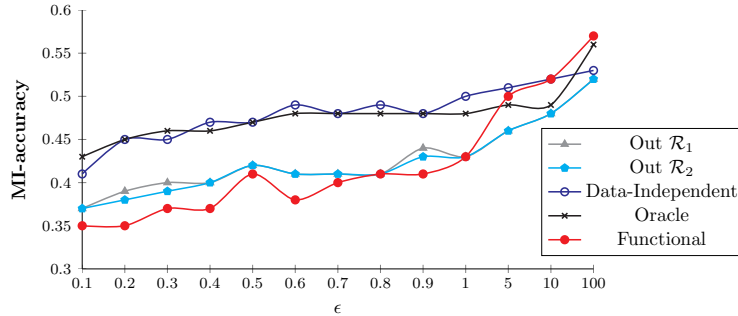


Figure 5: MI attack: Output Perturbation vs. Functional Mechanism. The x-axis is ε . The y-axis is the accuracy of MI attack. MI attacks become more effective for all three variants. For Oracle Output Perturbation and Functional Mechanism at $\varepsilon = .2$, the MI attack accuracy for the oracle mechanism is 45%, while is only 35% for the functional mechanism. Data-Independent and Oracle Output Perturbation are similar.

Figure 6 further demonstrates that, similar to their model accuracy, Data-Independent and Oracle Output Perturbation have very similar behavior on model invertibility (note that they both provide similar model accuracy that is better than other methods).

Comparison with Other Private Mechanisms. For MI we also compare our method with projected-histogram algorithm and objective perturbation. Specifically, We notice that in the previous work of Fredrikson et al. [11], they have demonstrated that projected-histogram algorithm actually leaks more information and produces models with higher MI accuracy than the functional mechanism (see the discussion under the

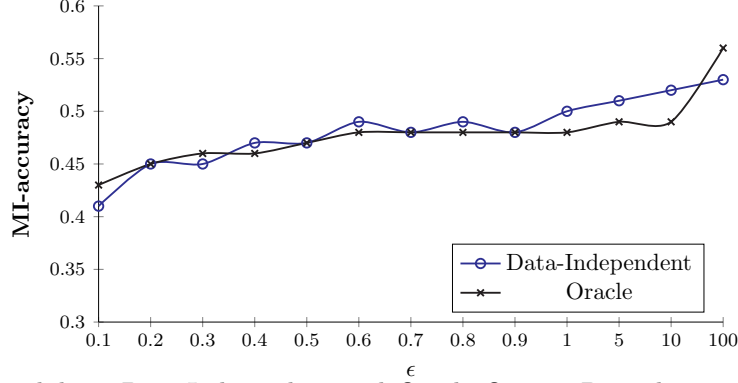
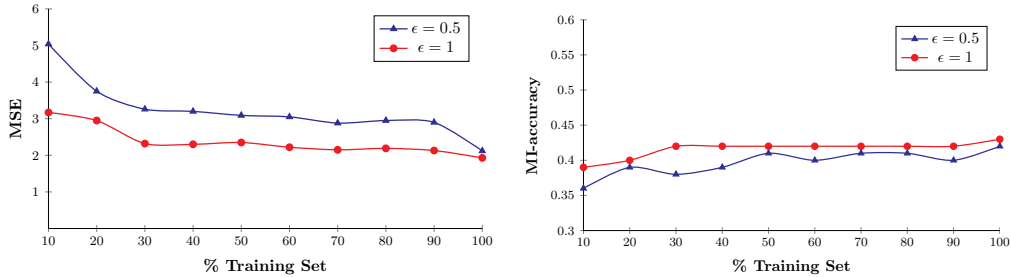


Figure 6: Model Invertibility: Data-Independent and Oracle Output Perturbation. The model inversion accuracy of these two algorithms are also very close with each other.

head “**Private Histograms vs. Linear Regression**” in Section 4 of their paper). For Objective Perturbation we find again that its MI accuracy is almost the same as our Data-Independent and Oracle Output Perturbation. This is not surprising because they have very close model accuracy.

For a fixed mechanism, better utility gives more effective MI attacks. For a fixed mechanism, we demonstrate that MI attacks get more effective as utility increases with the availability of more training data. Figure 7 shows the results for Tuned Output Perturbation. For $\epsilon_p = 0.5$, the mean square error drops



(a) Model Accuracy

(b) Model Invertibility

Figure 7: Privately-Tuned Output Perturbation with increasingly larger training sets. The x-axis is the size of training chunks, and the y-axis is MSE. We use the following experimental method: The entire data set gets randomly permuted in the beginning. The data is split again into 25 chunks, with the first 24 for training and the last one for validation. For each of the training chunks, a fraction of the training items is sampled at rate r . We then train the model using the data with Tuned Output Perturbation and evaluate their utility and MI attack accuracy. This experiment is repeated 100 times for $r = 10\%, 20\%, \dots, 100\%$.

from 5.05 (with $r = 10\%$ of available training data) to 2.9 ($r = 90\%$), while the MI attack accuracy increases from 36% ($r = 10\%$) to 40% ($r = 90\%$).

A Theoretical Analysis. At first sight the above two phenomena may seem somewhat peculiar. If MI attack is considered as a privacy concern, then we have empirically observed that some privacy concern becomes “worse”, while differential privacy gets “better.”

There is no contradiction here. Indeed, it suffices to observe that DP is a property of the learning “process”, while MI attack is on the “result” of the process. It is thus valid that the process satisfies a strong privacy guarantee, while the result has some other concerns. In the following, we give a “lower bound” result, which shows that, as long as the optimal solution of the learning problem is susceptible to MI attacks, improving DP/utility tradeoff will give effective MI attacks “eventually.”

The intuition is as follows: Suppose that MI attack will be effective at a hypothesis w^* such that

$L_{\mathcal{D}}(w^*) = L_{\mathcal{D}}^*$. That is, the MI attack is effective at a hypothesis we want to converge to. Now, suppose that the effectiveness of MI attack grows “monotonically” as we converge to w^* . Then, as long as the result of a learning algorithm converges to w^* , it will gradually give more effective MI attacks. We now give more details of this argument.

Assumptions. We make the following three assumptions on learning and MI attack: (i) The utility of a hypothesis is measured by its generalization error. (ii) Suppose that for w^* , $L_{\mathcal{D}}(w^*) = L_{\mathcal{D}}^*$, and MI attack is effective for w^* . (iii) As $|L_{\mathcal{D}}(w) - L_{\mathcal{D}}^*|$ gets smaller, the MI attack for w becomes more effective.

These assumptions are natural. For (i), almost all previous work measures utility this way. For (ii), since the ultimate goal of learning is to converge to the best possible hypothesis, assuming MI attack will be effective for such w^* is natural. Finally, closeness in $L_{\mathcal{D}}$ indicates that w and w^* are close in terms of their “functioning as a model.” Thus (iii) holds intuitively.

Better DP Mechanisms, More Effective MI attack. For any n , let S_n denote a training set of size n . Suppose that A' is an ε_p -differentially private mechanism with better privacy-utility tradeoff than the output perturbation mechanism A . Thus with high probability for $w \sim A'(S_n)$, $L_{\mathcal{D}}(w)$ is closer to $L_{\mathcal{D}}^*$ than that of w sampled from $A(S_n)$. Combined with (ii) and (iii), we have that MI attack is more effective for $w \sim A'(S_n)$.

One may note that the mechanism A' could be any differentially private mechanism, as long as it has better DP-utility tradeoff than *output perturbation*. For example, one can use the objective perturbation mechanism in [5] for generalized linear models, or exponential sampling based mechanisms in [3] if minimizing training error is the goal.

For a Fixed Mechanism, Better DP-Utility tradeoff gives More Effective MI attacks.. Theorem 14 and our discussion at the end of Section 3.2 have indicated that as n tends to the infinity, the amount of noise injected for ε_p -differential privacy vanishes to zero. Further, Theorem 20 and 21 imply that for any convex Lipschitz learning problem, the *output perturbation* mechanism converges to the optimal hypothesis w^* . Thus we are in the situation that as n increases, the output model has less noise yet it is closer to w^* . Therefore, by assumptions (ii) and (iii) MI attack is more effective for larger n .

A Bayesian View Point. In a paper by Kasiviswanathan and Smith [17], the authors give a Bayesian interpretation of the semantics of differential privacy. Informally speaking, given an arbitrary prior distribution over a collection of databases, what differential privacy guarantees is that *two posteriors* obtained in two worlds of neighboring databases are indistinguishable with each other.

What about the difference between *prior and posterior*? The same paper [17], and indeed the original paper by Dwork and Naor[9], have pointed out that it is impossible to bound the difference between prior and posterior under *arbitrary background knowledge*. Essentially, as long as the published information is “useful,” there exists some background knowledge that allows an adversary to learn and significantly modify his/her prior.

Unfortunately, in MI attacks, some moderate background information allows significant change of one’s prior. Therefore, while worsening MI attack is certainly not what one intended, it is also of no surprise that better differentially private mechanisms do not give better resilience against MI attack.

6 Related Work

Differential privacy was proposed in the seminal work of Dwork, McSherry, Nissim and Smith [8] and has become the de-facto standard for privacy. Our setting – learning a model differentially privately – was initiated in the work by Chaudhuri, Monteleoni and Sarwate [5]. Since the work of [5], a large body of work [3, 5, 7, 15] has been devoted to this line, culminating in a recent result by Bassily, Smith and Thakurta [3], which obtains tight error bounds for general convex-Lipschitz learning problems and generalized linear models. To this end, our work makes the underlying theme behind these works more explicit, namely, the connection between differential privacy and stability theory in machine learning.

On the other hand, the application of these results in practice seems slow-paced. The recent work by

Fredrikson et al. [11] highlights two unfortunate facts on training a differentially private pharmacogenetic model using the functional mechanism proposed by Zhang et al. [33]. The first issue is the low model accuracy even for weak differential privacy guarantees. For this problem, we show, both theoretically and empirically, that the simple output perturbation can provide a much better DP-utility tradeoff than the functional mechanism.

The second issue is about effectiveness of MI attacks. MI attack is a new kind of privacy attack that was first described in the same paper by Fredrikson et al. [11], where the authors demonstrate that a DP mechanism can only prevent MI attack with small privacy parameters. We show that the privacy concern of MI attack is essentially orthogonal to the concern of differential privacy. To the best of our knowledge, this is the first work that establishes such a connection.

7 Conclusion

In this paper, we considered issues raised by Fredrikson et al. about training differentially private regression models using functional mechanism. Through an explicit connection between differential privacy and stable learning theory, we gave a straightforward analysis showing that output perturbation can be made to obtain, both theoretically and empirically, substantially better privacy/utility tradeoff than the functional mechanism. Since output mechanism is simple to implement, this indicates that our method is potentially widely applicable in practice. We went on to apply our theory to the same data set as used by Fredrikson et al., and the empirical results are encouraging.

We also studied model inversion attack, a privacy attack raised in the same paper by Fredrikson et al. We observed empirically that better differentially private mechanisms lead to more effective model-inversion attacks. We analyzed theoretically why this is the case. We regard this result as a warning that care may be taken when learning things from a data set, even when strong differential privacy is guaranteed.

References

- [1] Y. Aono, T. H. L. T. Phong, and L. Wang. Fast and secure linear regression and biometric authentication with security update. 2015.
- [2] D. Applegate and R. Kannan. Sampling and integration of near log-concave functions. In *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing, May 5-8, 1991, New Orleans, Louisiana, USA*, pages 156–163, 1991.
- [3] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pages 464–473, 2014.
- [4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.
- [5] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.
- [6] F. J. Diaz and J. Yeh, Hung-Wen de Leon. Role of statistical random-effects linear models in personalized medicine. *Current Pharmacogenomics and Personalized Medicine*, 10(1):22–32, 2012.
- [7] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 429–438, 2013.
- [8] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.
- [9] C. Dwork and M. Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1):8, 2008.
- [10] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

- [11] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, August 20-22, 2014.*, pages 17–32, 2014.
- [12] M. Hardt. Towards practicing differential privacy. <http://blog.mrtz.org/2015/03/13/practicing-differential-privacy.html>, 2014.
- [13] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992.
- [14] International Warfarin Pharmacogenetic Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.
- [15] P. Jain and A. Thakurta. Differentially private learning with kernels. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 118–126, 2013.
- [16] Z. Jorgensen, T. Yu, and G. Cormode. Conservative or liberal? personalized differential privacy. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 1023–1034, 2015.
- [17] S. P. Kasiviswanathan and A. Smith. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR*, abs/0803.3946, 2008.
- [18] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *SIGMOD Conference*, pages 193–204, 2011.
- [19] N. Li, W. H. Qardaji, D. Su, and J. Cao. Privbasis: Frequent itemset mining with differential privacy. *PVLDB*, 5(11):1340–1351, 2012.
- [20] Y. Lindell and E. Omri. A practical application of differential privacy to personalized online advertising. *IACR Cryptology ePrint Archive*, 2011:152, 2011.
- [21] I. McKeague and M. Qian. Sparse functional linear regression with applications to personalized medicine. In F. Ferraty, editor, *Recent Advances in Functional Data Analysis and Related Topics*, Contributions to Statistics, pages 213–218. Physica-Verlag HD, 2011.
- [22] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, pages 94–103, 2007.
- [23] J. Reed, A. J. Aviv, D. Wagner, A. Haeberlen, B. C. Pierce, and J. M. Smith. Differential privacy for collaborative security. In *Proceedings of the Third European Workshop on System Security, EUROSEC 2010, Paris, France, April 13, 2010*, pages 1–7, 2010.
- [24] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
- [25] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.
- [26] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.
- [27] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.
- [28] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.
- [29] Y. Wang, C. Si, and X. Wu. Regression model fitting under differential privacy and model inversion attack. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 1003–1009, 2015.
- [30] M. Winslett, Y. Yang, and Z. Zhang. Demonstration of damson: Differential privacy for analysis of large data. In *18th IEEE International Conference on Parallel and Distributed Systems, ICPADS 2012, Singapore, December 17-19, 2012*, pages 840–844, 2012.
- [31] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett. Differentially private histogram publication. *VLDB J.*, 22(6):797–822, 2013.
- [32] J. Zhang, X. Xiao, Y. Yang, Z. Zhang, and M. Winslett. Privgene: differentially private model fitting using genetic algorithms. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013*, pages 665–676, 2013.
- [33] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett. Functional mechanism: Regression analysis under differential privacy. *PVLDB*, 5(11):1364–1375, 2012.

A Proofs

A.1 Proof of Proposition 10

Proof. Let $f_S(w)$ be the probability density function of $A(S)$. Due to ε -differential privacy, then for any $S, i, z', f_S(w) \leq e^\varepsilon f_{S^{(i)}}(w)$. Therefore

$$\begin{aligned}
& \left| \mathbb{E}[\ell(\tilde{A}(S), \bar{z})] - \mathbb{E}[\ell(\tilde{A}(S^{(i)}), \bar{z})] \right| \\
&= \left| \int \ell(w, \bar{z}) (f_S(w) - f_{S^{(i)}}(w)) dw \right| \\
&\leq \int |\ell(w, \bar{z})| |f_S(w) - f_{S^{(i)}}(w)| dw \\
&\leq B \int |f_S(w) - f_{S^{(i)}}(w)| dw
\end{aligned} \tag{1}$$

Note that $(e^{-\varepsilon} - 1)f_{S^{(i)}}(w) \leq f_S(w) - f_{S^{(i)}}(w) \leq (e^\varepsilon - 1)f_{S^{(i)}}(w)$, so $|f_S(w) - f_{S^{(i)}}(w)| \leq \max\{1 - e^{-\varepsilon}, e^\varepsilon - 1\}f_{S^{(i)}}(w)$. Plugging into (1) gives the first claimed inequality. The second inequality follows from the observation that $e^\varepsilon + e^{-\varepsilon} \geq 2$. \square

A.2 Proof of Lemma 12

Proof. By the definition of ϑ_S ,

$$\begin{aligned}
& \vartheta_S(u) - \vartheta_S(v) \\
&= \left(L_{S^{(i)}}(u) + \varrho(u) - \frac{1}{n}\ell(u, z') + \frac{1}{n}\ell(u, z_i) \right) \\
&\quad - \left(L_{S^{(i)}}(v) + \varrho(v) - \frac{1}{n}\ell(v, z') + \frac{1}{n}\ell(v, z_i) \right) \\
&= \vartheta_{S^{(i)}}(u) - \vartheta_{S^{(i)}}(v) + \frac{\ell(v, z') - \ell(u, z')}{n} + \frac{\ell(u, z_i) - \ell(v, z_i)}{n}.
\end{aligned}$$

Since u minimizes $\vartheta_{S^{(i)}}$ so $\vartheta_{S^{(i)}}(u) - \vartheta_{S^{(i)}}(v) \leq 0$, so

$$\vartheta_S(u) - \vartheta_S(v) \leq \frac{\ell(v, z') - \ell(u, z')}{n} + \frac{\ell(u, z_i) - \ell(v, z_i)}{n}$$

completing the proof. \square

A.3 Proof of Lemma 13

Proof. We have that for any $\alpha \in (0, 1)$,

$$\begin{aligned}
\vartheta_S(v) &\leq \vartheta_S(\alpha v + (1 - \alpha)u) \\
&\leq \alpha \vartheta_S(v) + (1 - \alpha) \vartheta_S(u) - \frac{\lambda}{2} \alpha (1 - \alpha) \|v - u\|^2
\end{aligned}$$

where the first inequality is because v is the minimizer of ϑ_S and the second inequality is by the definition of λ -strong convexity. By elementary algebra, this give that $\frac{\lambda}{2} \alpha \|v - u\|^2 \leq \vartheta_S(u) - \vartheta_S(v)$. Tending α to 1 gives the claim. \square

A.4 Proof of Theorem 14

Proof. Let $u = A(S^{(i)})$, $v = A(S)$. By Lemma 13,

$$\frac{\lambda}{2} \|u - v\|^2 \leq \vartheta_S(u) - \vartheta_S(v) \quad (1)$$

By Lemma 12,

$$\vartheta_S(u) - \vartheta_S(v) \leq \frac{\ell(v, z') - \ell(u, z')}{n} + \frac{\ell(u, z_i) - \ell(v, z_i)}{n} \quad (2)$$

Now because $\ell(\cdot, z)$ is ρ -Lipschitz, we have that $\ell(v, z') - \ell(u, z') \leq \rho \|v - u\|$, and $\ell(u, z_i) - \ell(v, z_i) \leq \rho \|u - v\|$. Plugging these two inequalities to (2), we have that $\vartheta_S(u) - \vartheta_S(v) \leq \frac{2\rho}{n} \|u - v\|$. Plugging this to (1) and rearranging completes the proof. \square

A.5 Proof of Lemma 17

Proof. We have

$$\begin{aligned} |L_{\mathcal{D}}(w) - L_{\mathcal{D}}(A(S))| &= \left| \mathbb{E}_{z \sim \mathcal{D}} [\ell(w, z) - \ell(A(S), z)] \right| \\ &\leq \mathbb{E}_{z \sim \mathcal{D}} [|\ell(w, z) - \ell(A(S), z)|] \end{aligned}$$

For every $z \in Z$, $\ell(\cdot, z)$ is ρ -Lipschitz, so $|\ell(w, z) - \ell(A(S), z)| \leq \rho \|w - A(S)\|_2$. Therefore

$$\mathbb{E}_{z \sim \mathcal{D}} [|\ell(w, z) - \ell(A(S), z)|] \leq \rho \|w - A(S)\|_2.$$

The proof is complete by observing that with probability at least $1 - \gamma$ over $w \sim \tilde{A}(S)$, $\|w - A(S)\|_2 \leq \kappa(n, \gamma)$. \square

A.6 Proof of Theorem 18

Proof. For any $S \sim \mathcal{D}^n$, we have that $L_{\mathcal{D}}(w) - L_{\mathcal{D}}^* = (L_{\mathcal{D}}(w) - L_{\mathcal{D}}(A(S))) + (L_{\mathcal{D}}(A(S)) - L_{\mathcal{D}}^*)$. For $L_{\mathcal{D}}(A(S)) - L_{\mathcal{D}}^*$, we know that $\Pr_{S \sim \mathcal{D}^n} [L_{\mathcal{D}}(A(S)) - L_{\mathcal{D}}^* > \varepsilon(n, \delta)] < \delta$. Further, for every $S \sim \mathcal{D}^n$, from Lemma 17, $\Pr_{w \sim \tilde{A}(S)} [L_{\mathcal{D}}(w) - L_{\mathcal{D}}(A(S)) > \rho \kappa(n, \gamma)] < \gamma$. The proof is complete by a union bound. \square

A.7 Proof of Theorem 20

Proof. Let A denote the rule of empirical risk minimization, and \tilde{A} be its output-perturbation counterpart which ensures ε_p -differential privacy. Then by Theorem 14, and corollary 9, with probability at least $1 - \gamma$ over $w \sim \tilde{A}(S)$, $\|w - A(S)\|_2 \leq \frac{4d \ln(d/\gamma) \rho}{\lambda n \varepsilon_p}$.

Together with Theorem 19, it follows that with probability at least $1 - \delta' - \gamma$ over $S \sim \mathcal{D}^n$ and $w \sim \tilde{A}(S)$,

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}^* \leq \frac{4\rho^2}{\delta' \lambda n} + \frac{4d \ln(d/\gamma) \rho^2}{\lambda n \varepsilon_p}.$$

Put $\delta' = \gamma = \delta/2$, thus with probability at least $1 - \delta$,

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}^* \leq \frac{4\rho^2}{\lambda n} \left(\frac{2}{\delta} + \frac{4d \ln(2d/\delta)}{\varepsilon_p} \right).$$

Asymptotically this is $O\left(\frac{\rho^2 d \ln(d/\delta)}{\lambda n \delta \varepsilon_p}\right)$. The proof is complete. \square

A.8 Proof of Theorem 21

Proof. Let $\bar{\ell}$ be defined as $\bar{\ell}(w, z) = \ell(w, z) + \frac{\lambda}{2}\|w\|^2$. $\bar{\ell}$ is λ -strongly convex and $(\rho + \lambda R)$ -Lipschitz. Let \bar{L}_S and \bar{L}_D be the empirical loss and true loss functions with respect to $\bar{\ell}$. Note that for any $w \in \mathcal{H}$, $\bar{L}_D(w) = L_D(w) + \frac{\lambda}{2}\|w\|^2$.

By Theorem 20, there is an ε_p -differentially private mechanism \tilde{A} such that with probability $1 - \delta$ over $S \sim \mathcal{D}^n$ and $w \sim \tilde{A}(S)$,

$$\bar{L}_D(w) - \bar{L}_D^* \leq O\left(\frac{(\rho + \lambda R)^2 d \ln(d/\delta)}{\lambda n \delta \varepsilon_p}\right) \quad (1)$$

Let $\bar{w} \in \mathcal{H}$ such that $\bar{L}_D(\bar{w}) = \bar{L}_D^*$ and $w^* \in \mathcal{H}$ such that $L_D^* = L_D(w^*)$. Because \bar{w} is the minimizer of $\bar{L}_D(\cdot)$, so

$$L_D(w^*) + \frac{\lambda}{2}\|w^*\|^2 \geq L_D(\bar{w}) + \frac{\lambda}{2}\|\bar{w}\|^2 \quad (2)$$

Combining (1) and (2) we have

$$\begin{aligned} L_D(w) - L_D^* &\leq O\left(\frac{(\rho + \lambda R)^2 d \ln(d/\delta)}{\lambda n \delta \varepsilon_p}\right) + \frac{\lambda}{2}(\|w^*\|^2 - \|w\|^2) \\ &\leq O\left(\frac{(\rho + \lambda R)^2 d \ln(d/\delta)}{\lambda n \delta \varepsilon_p}\right) + \frac{\lambda R^2}{2} \\ &\leq O\left(\frac{2\rho R d \ln(d/\delta)}{n \delta \varepsilon_p} + \frac{\rho^2 d \ln(d/\delta)}{\lambda n \delta \varepsilon_p} + \lambda R^2\right) \end{aligned}$$

where the second inequality is because the hypothesis space is R -bounded. Putting $\lambda = \frac{\rho}{R} \sqrt{\frac{d \ln(d/\delta)}{n \delta \varepsilon_p}}$ gives the claimed bound. \square

A.9 Smooth Problems

The following lemma bounds training error for smooth learning problems. The lemma is first proved by Chaudhuri et al. [5] for generalized linear models.

Lemma 22. *Consider a learning problem (\mathcal{H}, Z, ℓ) . Suppose that ℓ is λ -strongly convex, ρ -Lipschitz, and β -smooth. Let A be ERM and \tilde{A} be its output perturbation variant as described in Algorithm 1 with privacy parameter $\varepsilon_p > 0$. Then for any training set $S \subseteq Z^n$, we have with probability at least $1 - \gamma$ over $w \sim \tilde{A}(S)$,*

$$L_S(w) - L_S(w^*) \leq \beta \left(\frac{4d \ln(d/\gamma) \rho}{\lambda n \varepsilon_p} \right)^2.$$

where $L_S(w^*) = \min_{w \in \mathcal{H}} L_S(w)$.

Proof. Using Mean Value Theorem, we have that for some $\alpha \in (0, 1)$, $L_S(w) - L_S(w^*) = \nabla L_S(\alpha w + (1 - \alpha)w^*)(w - w^*)$, which is upper bounded by $\|\nabla L_S(\alpha w + (1 - \alpha)w^*)\| \|w - w^*\|$ by the Cauchy-Schwarz inequality.

Note that w^* achieves the minimum, so $\nabla L_S(w^*) = 0$. Thus $\|\nabla L_S(\alpha w + (1 - \alpha)w^*)\| = \|\nabla L_S(\alpha w + (1 - \alpha)w^*) - \nabla L_S(w^*)\|$, which is upper bounded by $\beta \|\alpha(w - w^*)\| \leq \beta \|w - w^*\|$ because ∇L_S is β -Lipschitz. Therefore we have that $L_S(w) - L_S(w^*) \leq \beta \|w - w^*\|^2$.

Finally, note that with probability at least $1 - \gamma$ over $w \sim \tilde{A}(S)$, $\|w - w^*\| \leq \frac{4d \ln(d/\gamma) \rho}{\lambda n \varepsilon_p}$, so $L_S(w) - L_S(w^*) \leq \beta \left(\frac{4d \ln(d/\gamma) \rho}{\lambda n \varepsilon_p} \right)^2$. \square